SANDIA REPORT
SAND201X-XXXX
Unlimited Release
Printed September 2017

Exploiting Social Media Sensor Networks through Novel Data Fusion Techniques

Tina Kouri
Ali Pinar

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@osti.gov
Online ordering: http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.gov
Online order: https://classic.ntis.gov/help/order-methods/

SAND201X-XXXX
Printed November 2017
Unlimited Release

Exploiting Social Media Sensor Networks through Novel Data Fusion Techniques

Tina Kouri
Tasking Planning and Mission Management (6323)
Sandia National Laboratories
P. O. Box 5800
Albuquerque, New Mexico 87185-MS1243

Abstract

Unprecedented amounts of data are continuously being generated by sensors ("hard" data) and by humans ("soft" data), and this data needs to be exploited to its full potential. The first step in exploiting this data is determine how the hard and soft data are related to each other. In this project we fuse hard and soft data, using the attributes of each (e.g., time and space), to gain more information about interesting events. Next, we attempt to use social networking textual data to predict the present (i.e., predict that an interesting event is occurring and details about the event) using data mining, machine learning, natural language processing, and text analysis techniques.

This page intentionally left blank

TABLE OF CONTENTS

FIGURES

TABLES

This page intentionally left blank

NOMENCLATURE

Abbreviation	Definition
ANSS	Advanced National Seismic System
DF	Data Fusion
MDRC	Modified Dempster's Rule of Combination
NLP	Natural Language Processing
NSMP	National Strong Motion Project
USGS	United States Geological Survey

This page intentionally left blank

1. INTRODUCTION

National security is essential for the well-being of the United States, but it is becoming more difficult to ensure security as the threats against the United States continue to increase. An important tool for national security is data utilization since data is being generated at unprecedented rates[1] from a variety of sources. According to [2], the intelligence world is changing due to changes in the world where available data is "dispersed, not concentrated" and "open to several sources". The authors of [2] also state that community needs to be "open to strong utilization of new information technologies to take profit of the information (often contradictory) explosion (information density doubles every 24 months and its costs are halved every 18 months [3])". It is imperative that we use every available source of information to protect our nation.

Sources of data include traditional sensors (e.g., space and ground sensors) and nontraditional sensors (e.g., human generated social media text). Traditional sensors are an essential source of data since they provide unbiased and trusted data points for current events, but they may not be sufficient for making operational decisions due to limited field-of-views or visual obstructions [4]. Non-traditional data sources are useful for gathering exclusive information because of their openness and their abundance of information, but may be ineffective due to their unreliability [4]. Social media platforms (e.g., Facebook and Twitter) encourage users to post anything from anywhere at any time [5], essentially creating citizen journalists [6].

Although there have been several significant advances in big data research and in utilizing the available social media data, there is still a lot of work to be done to fully exploit the knowledge contained in the vast amounts of data. [7]

A first step in fully exploiting the information contained in multiple data sources is Data Fusion (DF). DF algorithms must determine which data, from the various sources, are related to each other. For example, we may wish to determine which Twitter[2] posts (i.e., tweets) are associated with a specific earthquake that is reported by the USGS Quake ML service[3] . Section 2 describes approaches to DF.

In addition to fusing data in order to gain more information about an event from social media, we may want to use social media to predict a current event. For example, it is useful to predict that an earthquake is occurring using tweets since users typically start posting about an earthquake within seconds. USGS researchers have found that, for some earthquakes, information about the earthquakes is available from tweets before their equipment is able to publish information about the earthquake [8]. The few minutes warning may help people get to safe location. Section 3 describes approaches to predicting the present.

[1] "90%" of the data in the world today has been created in the last two years [1]
[2] http://www.twitter.com
[3] https://earthquake.usgs.gov/earthquakes/feed/v1.0/quakeml.php

This page intentionally left blank

2. DATA FUSION

Vast amounts of data are being generated from many different sources and many of those sources are referring to the same event. For example, USGS reports various statistics (e.g., location, magnitude, and time) for earthquakes it senses and Twitter users tweet about their earthquake experiences. Each of the sources tell a different part of the same story using their own unique properties and attributes. DF is designed to merge the data for an interesting event in order to gain more information about the event.

The DF problem is formally defined in Definition 2.1 and an example instance is shown in Example 2.1.

Definition 2.1. Given:

- A "hard" data set of interesting events
- A "soft" data set of text posts, where each text post has at least one of the following properties
    - Post is from a compatible time period
    - Post is from a compatible geographic region
    - Post mentions the geographic region
    - Post mentions the interesting event

Use attributes of each event/post to fuse the "hard" and "soft" data. Extract information from the "soft" data in order to gain more information about the observed interesting event.

Example 2.1. Use twitter to gain more information about holiday flight delays to/from Albuquerque

- Our set of interesting events is the hard sensor observations of delayed flights over the 2015 Christmas holiday season
- Our soft data set is all tweets. We are interested in posts that
    - Originate late December or early January
    - Originate in Albuquerque or nearby
    - Mention Albuquerque
    - Mention things related to flight delays

Note that Definition 2.1 limits the amount of textual data we consider due to the vast amounts of textual data that are available.

2.1. Related Work

There has been a lot of research in exploiting information from big data sources [9], but the next big challenge in big data research is DF [7,10–12]. Some of the current

approaches to the DF problem rely on user input and use traditional data mining techniques [11,13] or mathematical techniques. Mathematical techniques include Modified Dempster's Rule of Combination (MDRC) [4] and random set theory techniques [14]. Researchers have identified many challenges with developing DF algorithms, which include performing the fusion and analysis of the results in real-time [9].

In [13], the authors survey several recent DF algorithms and compare the results on flight and stock market data sets. They claim that many DF algorithms rely on some mechanism of voting which may or may not take into account the trustworthiness of the source or whether or not some sources copied from each other (e.g., a Twitter retweet). They found that many of the algorithms had to tradeoff between efficiency and precision, which implied that the best algorithm to use was dependent on the data set being analyzed and the application the results were used in. The authors found that the approaches which considered the trustworthiness of the source were the most promising.

2.2. Algorithm

We have developed an algorithm to fuse "hard" and "soft" data based on the attributes associated with each. The algorithm fuses all data that are within the specified bounds for each attribute evaluated. Each attribute computation is normalized by dividing by the maximum difference between any two potentially fused events. The attributes are weighted and summed together to get an overall score for the fusion. See Example 2.2 for an illustration of how score are computed. Fused data is recommended using a nearest-neighbor algorithm.

Example 2.2. Let x be an interesting event that occurs at position l_x and time t_x. Let y be a potentially fused event that occurs at position l_y and time t_y. Let L be the (absolute) maximum distance between x and any other potentially fused event. Let T be the (absolute) maximum time between x and any other potentially fused event. When the time and position attributes are equally weighted, the score for y is

$$\frac{1}{2}\left(\frac{|l_x - l_y|}{L}\right) + \frac{1}{2}\left(\frac{|t_x - t_y|}{T}\right) \tag{1}$$

In Example 2.2 we used time and location attributes to illustrate how we fused hard and soft data based on attributes. Although there are other attributes that we may use (e.g., sentiment), we have found that time and space are the best indicators of related events, which is a logical and expected result. According to [15] the first law of geography which states "everything is related to everything else, but near things are more related than distant things" [16]. The same principle applies to time since things that occur close in time are usually more related than things that occur at significantly different times. Additional attributes may be checked by the algorithm, which may be useful depending on the data set and the DF application.

Figure 2.1. Tweets following an earthquake in CA

When considering time and space of textual data, it is possible that more than one time or location may be referenced. Therefore, it is imperative that the algorithm be able to account for multiple times and locations. In our algorithm, we use the point from the soft data which is closest to the hard data point we are considering.

It is also possible that textual does not have a location or relevant time. For example, during an earthquake users typically post a short statement which simply indicates that there is an earthquake, e.g., the entire tweet is simply "earthquake". Figure 2.1 shows some example tweets following an earthquake in California. If a potentially fused event does not have an attribute then that attribute is ignored when computing the score and the weights are adjusted accordingly.

Since we are working with user-generated text data, it is common to encounter text with a date that is not complete (e.g., December 2016) or a location that is not precise (e.g., Albuquerque, NM). The algorithm completes the date and adds some uncertainty to the date (e.g., +/- 1 day) or location (e.g., +/- 1km). The amount of uncertainty added depends on how uncertain the data is. The differences between two uncertain attributes is computed using the point where their ranges are closest together.

In user generated text data, it is possible that an attribute is referred to more than once. For example, text may state something such as "Last Tuesday I did something, but Friday I did something else". The algorithm allows objects to have more than one of each attribute.

Natural Language Processing Since DF is dependent on the attributes of the data we must utilize the appropriate tools to extract information from textual data. Some of the data is easily obtained using the meta data (e.g., a time stamp), but much of the interesting data will come from the text itself. For example, a twitter user may have their location set to California, but post about an event they are at in another state.

Several tools have been developed to process text data and extract information from the text [17–20]. The tools are able to extract dates and times [18], locations [18], sentiment [18], topic [17, 19, 20], and type of speech (e.g., question or statement) [19]. Researchers are continuing to improve the Natural Language Processing (NLP) algorithms to more efficiently and effectively extract useful information from text data. In this work, we use the Stanford CoreNLP library [18] to extract dates, locations, and sentiment from text data.

2.3. Software

We have developed DF software, using the Java programming language, to test and visualize the algorithms developed for DF. This software is designed such that any DF algorithm may be used so long as it implements the specified, simple, interface. The

software provides tools for recommending fused data, using user specified relationships between data items, and visualizing fused data.

The interface for the DF algorithms computes a score for each data item which is fused with the selected event. This score may be used to recommend fused data results which are the most relevant. We have implemented a nearest-neighbor recommender system, but the software is designed to allow developers to plug-in their own recommender systems. The software is also designed to allow recommender systems to utilize user-defined ratings for fused data items.

Users can define related data (e.g., user can define that Sandia NM is related to Sandia CA) which is useful learn more about an event which is not occurring in a single geographic location. For example, if both Sandia NM and Sandia CA are having an HBE event then, knowing that the two locations are related and seeing that they both have text data related to that event, we could glean additional information about the event and the scope of the event. Relevant locations, like relevant content in text, cannot be linked without external knowledge [9]. For example, Magic Mountain and Six Flags California refer to the same location, but they would not be linked without additional information.

The software also provides tools for visualizing fused data, which includes plotting the fused data on a map, graphing trends over time, and finding trending keywords in fused textual data.

It is often useful to see the results of a DF algorithm. We display the results of DF on a Java WorldWind[4] map . The data item which represents the interesting event is displayed as a point on the map and each fused data item is displayed as a point on the map in a different color. Each of the points which are placed on the map are color-coded based on their data type (e.g., “hard” sensor data is blue, “soft” textual data is green, location data is red).

Each fused data item is selectable to allow users to find additional fused data items. For example, if we are detecting tweets which are associated in time and space with a specific earthquake then we may wish to find which other earthquakes could possibly be related to one of the fused tweets.

Users may also to wish to create charts to visualize trends over time. For example, create a chart which shows the number of text messages and sensor events over time. The software uses JFreeChart[5] to generate graphs requested by the user. Users may select the time period which should be displayed on the chart as well as the periodicity (e.g., days, weeks, months).

Fused textual data may provide additional information about an event based on which words are trending. Therefore, the software provides functionality to determine trending keywords. This functionality is implemented using the Foundry library [21]. The user interface for finding trending terms includes several, optionally enabled, filters and tools, which include: (1) a stop words filter so that common words such

[4] https://worldwind.arc.nasa.gov
[5] http://www.jfree.org/index.html

'the' do not appear on the list of trending terms, (2) a minimum (and maximum) word length filter so that excessively small words do not appear on the list of trending terms, (3) a Porter English Stemmer so that words with the same root (e.g., shaking and shake) are considered the same word, and (4) a synonym substitution tool so that similarly defined terms (e.g., earthquakes and quakes) are grouped together.

The software has several additional features, including: (1) Filters which allow users to cull data as it is being input from a file so that only relevant data is considered, (2) Attempt to determine the type of data held in each column of a CSV file based on its column name, (3) Users may select a maximum number of fusion results to display which limits the results to those which are most relevant, and (4) Users may add additional dates and locations to events.

2.4. Future Work

In Section 2.3, we stated that software is designed to allow a recommender system to utilize user-defined rankings of fused data items in order to recommend future fused data items. It would be useful to develop and implement a more advanced recommender system which uses this data and other additional relevant data.

Due to the nature of human-generated textual data, it is more error-prone than traditional sensor data. Users may post erroneous information for a variety of reasons (accidental and purposeful) which means that the trustworthiness of data sources must be evaluated and incorporated in the DF and recommendation algorithms.

This page intentionally left blank

3. PREDICTION

Researchers have developed algorithms which use Twitter to predict and/or provide additional information about events which trend on social media [22]. We have attempted to use Twitter to detect and learn about important, yet smaller scale events. We selected as our exemplar problem detecting smaller earthquakes (yet still large enough to be felt by humans) or earthquakes which occur in areas which are not densely populated by Twitter users. We explore developing algorithms to "predict the present" (see also [23]), namely the time[6] , location, and magnitude of earthquakes using Twitter as a sensor.

We focused primarily on earthquakes which occurred in the United States whereas much of the previous research in this area focused on Japan. The previous research selected Japan since there are a large number of Twitter users in areas which frequently have earthquakes [24–26]. The authors of [26] state that the only area where the population of Twitter users truly overlaps with earthquake occurrences is Japan. They do acknowledge that there are areas of the United States (e.g., Los Angeles and San Francisco), Indonesia, Turkey, Iran, and Italy have some intersection, but their densities are significantly lower than Japan. We decided to focus on the United States, which also has a significant active Twitter user population and has earthquake activity (see Section 3.2), but the Twitter user population is not as dense as Japan and there are fewer significant earthquakes.

Section 3.1 described related work. Section 3.2 describes the data we use to explore this problem and some of the problems we encountered while working with this data. Section 3.3 describes our approach and ideas for solving this problem.

3.1. Related Work

Researchers have used various forms of social media, such as twitter [4,8,24–30], blogs [31–33], search trends (e.g., Google Trends[7]) [23,34], and news articles [35] to make predictions and to determine public opinion about events. They have created models to predict a variety of things, including elections [28], the stock market [29,35], product sales [27,31–34], threats [4,30], and natural disasters [8,24–26]. Keyword [4, 8, 24–26, 28] and sentiment [27–29, 32, 33] features are often used to create prediction models.

The authors of [36] discuss the Google flu trends algorithm, which was originally created in 2008. They discuss the errors in their model and why they significantly overestimated the number of flu occurrences in 2011 [37]. During the flu season of 2011, the algorithms of their model were affected by the increase queries due to increased media coverage. Part of their algorithm looks for sharp increases in queries which they labeled as "inorganic" and remove them from their model, but they failed to account for increases in queries over an entire flu season. Filtering out large spikes will not work for detecting earthquakes, since large spikes are expected whenever an earthquake occurs.

[6] Time is "predicted" as simply the earliest Tweet which is known to be associated with the earthquake
[7] https://trends.google.coml/trends/

3.2. Data

In this section we describe the data sources we considered for the prediction problem and our approach to acquiring data.

3.2.1. Twitter

Twitter is a social network microblogging service where users post short messages (i.e., microblogs), called Tweets [38]. Many active twitter users update their microblog several times per day and the updates are not necessarily unique (e.g., a user may retweet another user's post). Users post about anything and everything - see Figure 3.1 for a sample of Tweets from the "Spritzer" stream. Twitter also has an API which allows researchers to download Tweets along with metadata about the Tweets. Due to the abundance and variety of Twitter data, researchers have started developing algorithms to extract useful information from the service.

Figure 3.1. Tweets retrieved from the "Spritzer" stream on June 2, 2017

History and Statistics. The Twitter service started in 2006 [39] and has since become very popular among social network users. According to www.statista.com, as of the second quarter of 2017, there are approximately 328 million monthly active Twitter users worldwide [40]. Figure 3.2 shows the growth of the Twitter service since 2010.

The millions of Twitter users are dispersed throughout the world, but some areas of the world have significantly more users than other parts of the world. Figure 3.3 shows the number of Twitter users in 32 different countries, ranked by the percentage of Twitter users. Figure 3.4 is a heat map and a cluster map of the 1,000,000 tweets for an arbitrary day in 2017. Clearly some parts of the world have more active Twitter users than other parts of the world.

Analyzing Tweets. Text analysis is an essential step in extracting information from tweets and several techniques have been developed for text analysis. Bag-of-words

model textual analysis is simple to understand and implement, but researchers have found that results from bag-of-words model are often inferior to methods which consider named entities, noun phrases, or appraisal words [32,35]. The bag-of-words model considers all of the words in the text, but does not consider structure or context, whereas more advanced methods are designed to perform more in-depth analysis.

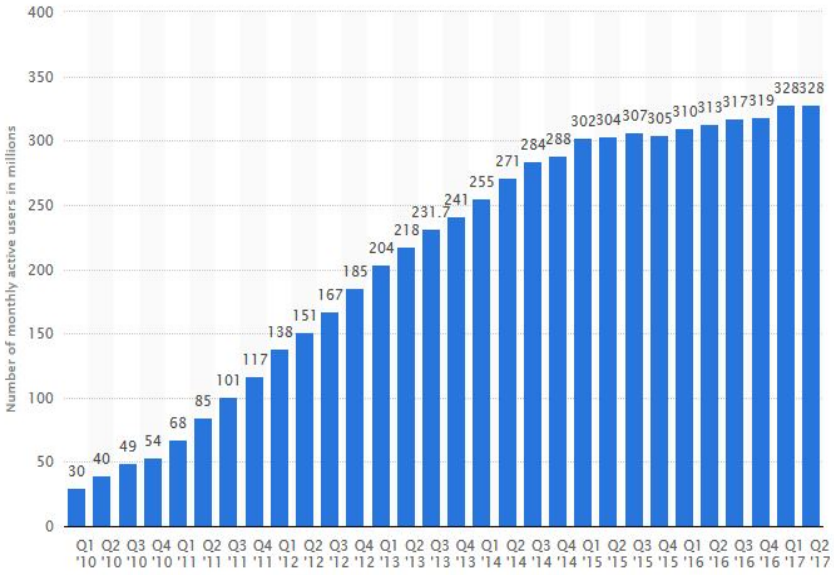


Figure 3.2. Number of active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2017 (in millions) [40]

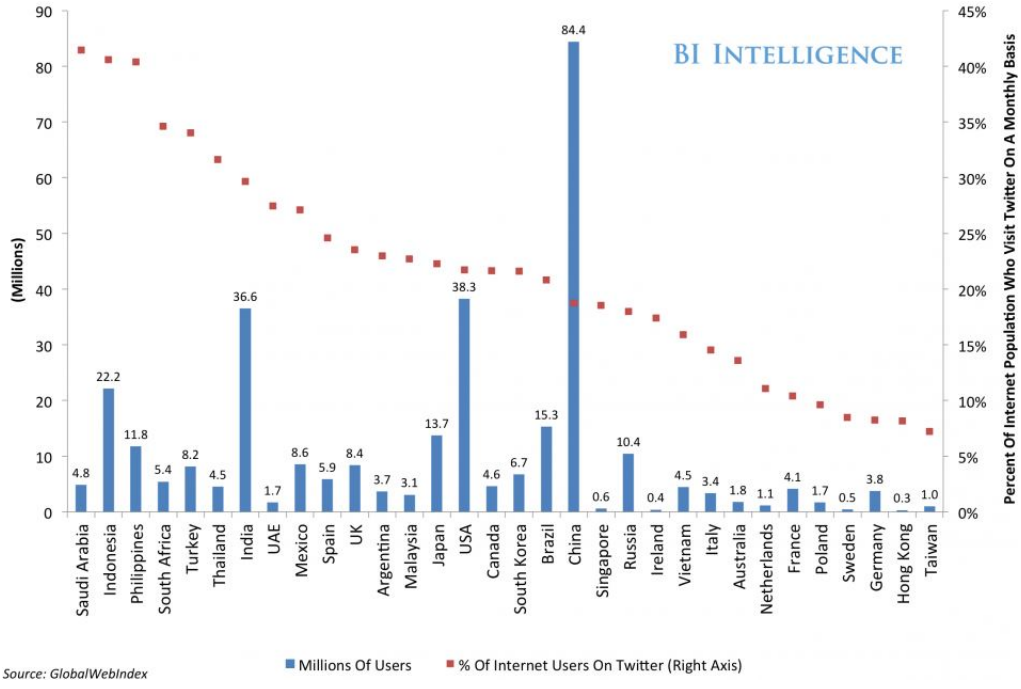
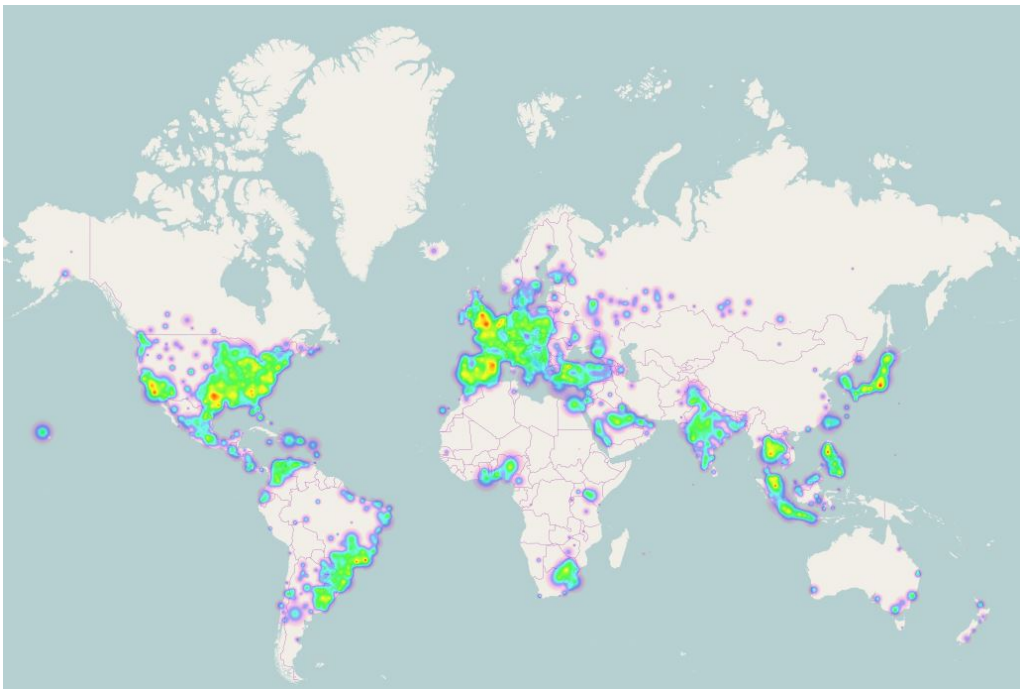
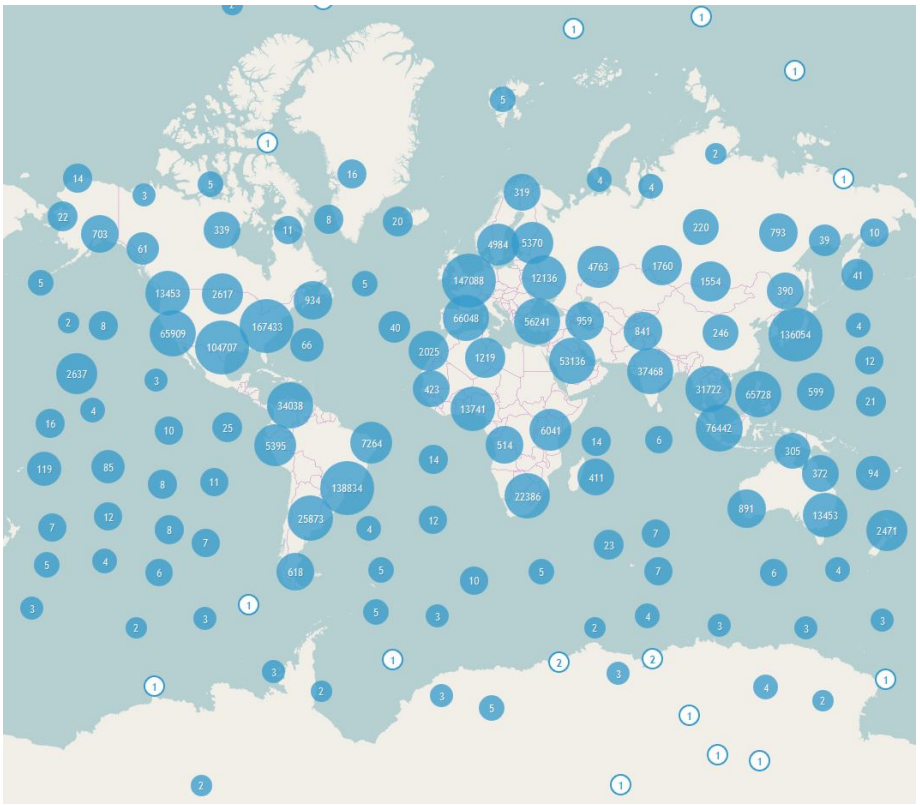


Figure 3.3. Countries ranked by percentage of Twitter users[41]



(a) Heat Map



(b) Cluster Map

Figure 3.4. The One Million Tweet Map [42]

Challenges. There are many unique challenges that arise when working with Twitter because of its nature. The authors of [9] describe some of Twitter's distinctive features: (1) time sensitivity (bloggers may make several updates each day which may or may not pertain to a recent event), (2) short length (tweets are limited to 140 characters), (3) unstructured phrases, and (4) abundant information.

We encountered many similar problems as we worked with the Twitter data, including:

1. The data is incomplete
   For example, many posts regarding an earthquake simply state "earthquake" without any reference to location or magnitude. It is possible for users to geo-tag their tweets, but very few users add this tag.

2. The data is noisy [37]
   Many of the tweets are not relevant to any interesting event (e.g., a user may post about their dinner) or a tweet may be posted long after an event (e.g., "that earthquake last week was very scary")

3. The data is unreliable
   There is no requirement that a tweet contain factual information. For example, a user may post about a non-existent earthquake just to see if they can get the topic trending. There is also no requirement that a user post about an event at all. For example, if there is an earthquake users may not post about it if they are more worried about their safety. The authors of [43] observe that popular sources of big data (e.g., twitter) are not "designed to produce valid and reliable data amenable for scientific analysis."

4. The data is unstructured
   Users may post about an event using any vocabulary or grammar they wish.

Trustworthiness. As mentioned previously, Twitter data is unreliable since it is not "designed to produce valid and reliable data amenable for scientific analysis" [43]. Therefore, current research initiatives are investigating how to determine whether social media information is true or false.

In [44], the authors develop a service[8] to determine if claims are true or false based on the footprint of how they propagate through social media. They compute a "spreading" score and a "skepticism" score which reflect how far a claim is spreading and a ratio of tweets doubting a claim, respectively. To determine if a Tweet expresses doubt in a claim, the authors use keywords which they found to be common in Tweets. The authors found that most of the time false claims do not spread in the same manner as true claims. They also observe that some false claims never gain enough traction, or spread far enough, that others express doubt in them.

In [5], the authors describe a rumour detection/classification system which gathers and analyzes the judgements of users to track the veracity status of a rumour as the

[8] http://twittertrails.com/

judgements are being made. The authors define a rumour as a "piece of circulating information whose veracity status is yet to verified at the time of posting".

3.2.2. Earthquakes

In the United States, the size of an earthquake is measured as magnitude of the earthquake. The lowest magnitude that may be felt by a human is just around magnitude 2.5, but it is rare for this level of earthquake (e.g., it may be felt by a few people on the upper floors of buildings). Magnitude 3.0 earthquakes will be noticeable by most people indoors. Table 3.1 shows the relationship between earthquake magnitude and the amount of energy released. [45]

USGS. The United States Geological Survey (USGS) was created in 1879 by congress and their mission is to "*provide science about the natural hazards that threaten lives and livelihoods, the water, energy, minerals, and other natural resources we rely on, the health of our ecosystems and environment, and the impacts of climate and land-use change*" [46]

One important natural hazards which is researched by USGS is earthquakes. They have developed several tools (e.g., QuakeML[9] , Advanced National Seismic System (ANSS)[10], National Strong Motion Project (NSMP)[11]) to help them better understand many aspects of earthquakes.

The QuakeML service reports earthquakes that are detected by USGS equipment and provides a REST API for users to access this information. Figure 3.5 shows earthquakes with magnitude 4.5 or higher and Figure 3.6 shows earthquakes with magnitude 3.0 to 4.5. Table 3.2 summarizes earthquakes reported by the QuakeML service.

Table 3.1. Earthquake Magnitude and Energy

Magnitude	Energy Release (Equivalent Pounds of Explosive)
10	120,000,000,000,000
9	4,000,000,000,000
8	120,000,000,000
7	4,000,000,000
6	120,000,000
5	4,000,000
4	120,000
3	4,000
2	120

[9] https://earthquake.usgs.gov/earthquakes/feed/v1.0/quakeml.php
[10] https://earthquake.usgs.gov/monitoring/anss/
[11] https://earthquake.usgs.gov/monitoring/nsmp

3.2.3. Acquiring Data

Social Media Data. There are two approaches which are used for sampling social media streams: (1) top-down approach and (2) bottom-up approach. In the top-down approach, social media is specifically sampled for posts related to a known event. This may be done by filtering in time and space for the specific event and by filtering the text for keywords[12] . Alternatively, the bottom-up approach samples the entire social media stream to retrieve posts related to real-time event or the posts from the stream are marked by a user to indicate if they are relevant for the topic being studied. [5] For this project we ended up using a bottom-up approach, in part due to the limitations of acquiring Twitter data.

We downloaded twitter data from the "spritzer" stream[13] for several months in 2017 using the Twitter4J API[14] and we received historical Twitter data from another researcher at Sandia National Labs for several days in 2016 which had earthquakes reported by USGS. Note that the Twitter API is written for users to retrieve real-time Tweets or to retrieve very recent Tweets. Therefore, it very challenging to acquire Tweets for past events. The Twitter data was put into two separate databases. The first database contained all of the tweets which we acquired. The second database only contained tweets which contained a keyword related to earthquakes in its text. We primarily use the keywords database for earthquake detection.

For this project we used a fixed set of keywords, but it is important to note that as a topic diffuses through social media, the terms used to describe that topic may change [47,48].

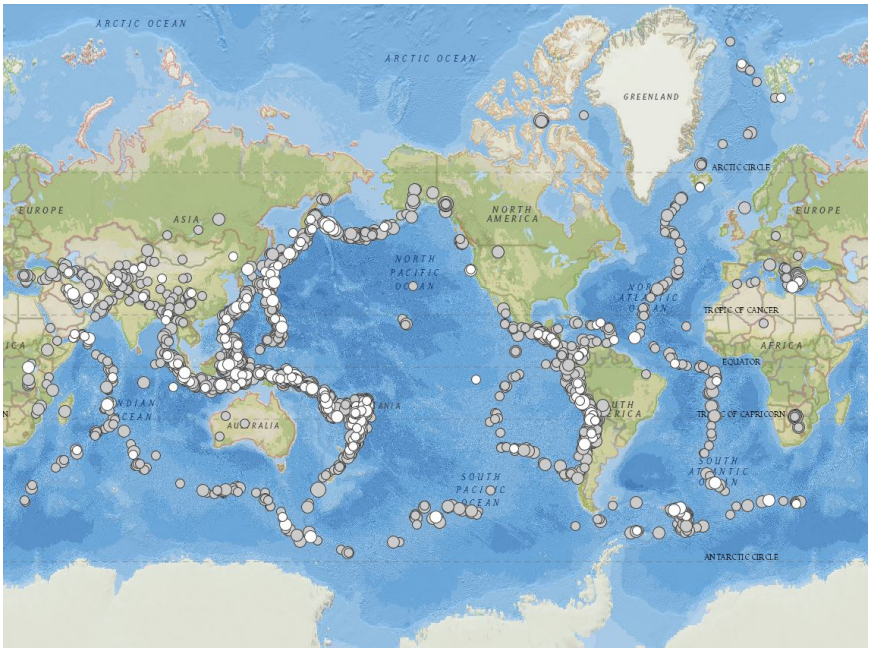
Overall, we gathered over 155 million tweets. Approximately 530,000 (0.3%) of the tweets gathered are geo-tagged. Approximately 1.6 million of all of the tweets have a keyword associated with earthquakes, of those tweets approximately 7000 (0.4%) are geo-tagged. When we attempted to limit Tweets messages to those within a specified geographic area for a given time period, the API often returned zero relevant Tweets.

Earthquake Data. We acquired earthquake data using the USGS QuakeML service, which reports earthquakes that the USGS detects and provides a REST API for users to retrieve information about the earthquakes. Figure 3.5 shows earthquakes with magnitude 4.5 or higher and Figure 3.6 shows earthquakes with magnitude 3.0 to 4.5. Table 3.2 summarizes earthquakes reported by the QuakeML service.

Notice that there are more earthquakes (both large and small earthquakes) in Japan than there are in the United States. Recall from Section 3.2.1 (Figure 3.2) that the density of internet users that are active Twitter users in Japan is larger than the in the United States.

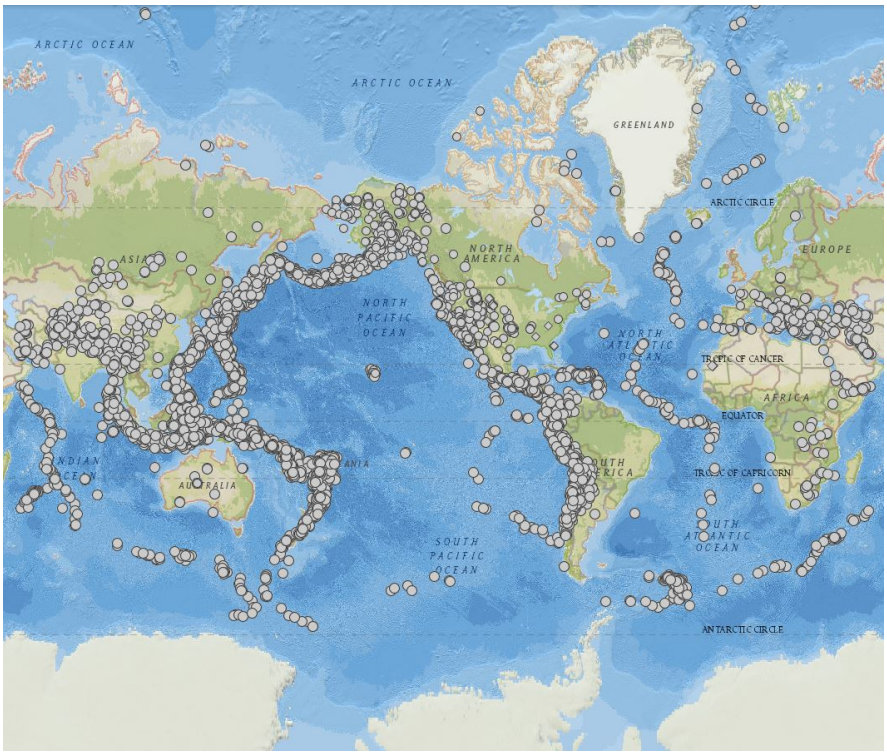
[12] Filtering textual data based on keywords is a commonly used technique when working with social media data [4]
[13] The "Spritzer" Twitter stream provides approximately 1% of all tweets
[14] http://twitter4j.org/en/

(a) January 1, 2016 – January 1, 2017

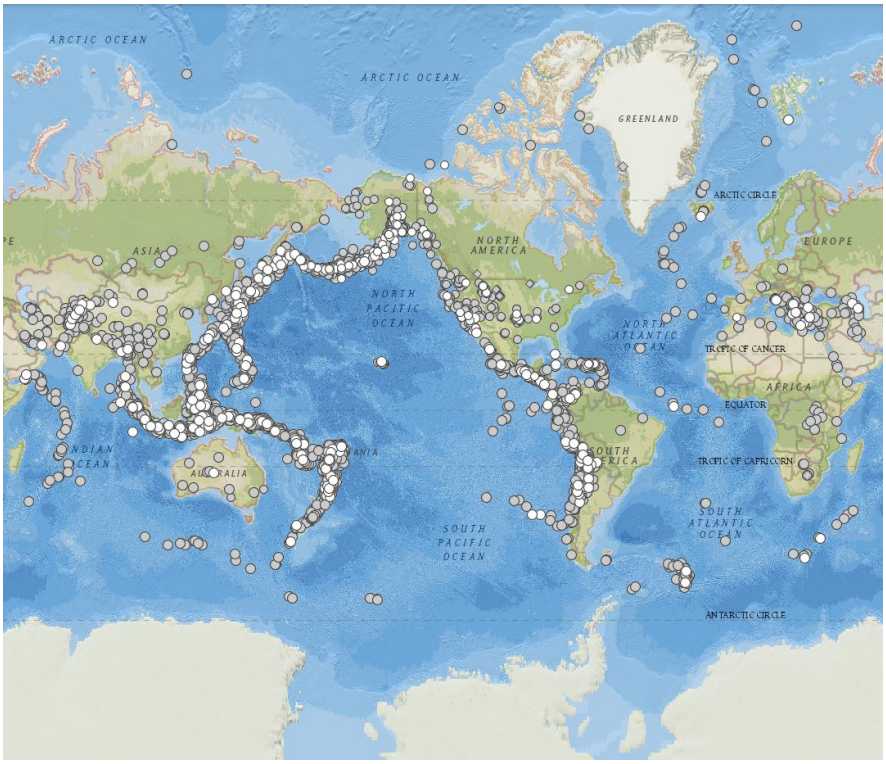


(b) January 1, 2017 – July 31, 2017

Figure 3.5. Earthquakes with magnitude 4.5 or higher [49]



(a) January 1, 2016 – January 1, 2017



(b) January 1, 2017 – July 31, 201

Figure 3.6. Earthquakes with magnitude 3.0 to 4.5 [49]

Table 3.2. Number of earthquakes reported by USGS

Magnitude	Number of Earthquakes			
	Jan 1, 2016 – Dec 31, 2016		Jan 1, 2017 – Jul 31, 2017	
	United States	Worldwide	United States	Worldwide
> 5.0	12	1664	5	860
4.5 - 5.0	10	5754	3	2404
4.0 - 4.5	65	8963	23	3732
3.5 - 4.0	243	1830	108	958
3.0 - 3.5	905	2869	345	1535
2.5 - 3.0	2582	6828	1083	3631
2.0 - 2.5	3390	11028	2017	6943

3.3. Learning from the Data

After collecting the data we compared the earthquake data from USGS with the Twitter data to look for correlation of known earthquakes. As expected, in populated areas with a large number of active twitter users the number of tweets which mention the earthquake dramatically increase after a significant earthquake - see Figure 3.7. Some examples of tweets related to the earthquake are shown in Figure 3.8.

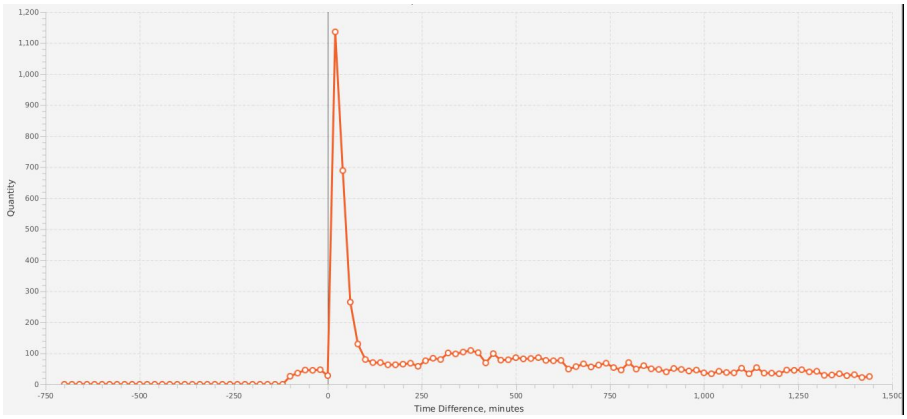


Figure 3.7. Number of tweets with earthquake keywords following a magnitude 5 earthquake in California in June of 2016

Figure 3.8. Tweets following a magnitude 5 earthquake in California in June of 2016

For smaller earthquakes, the data did not have nearly as much correlation

After collecting the data, we initially attempted to detect earthquakes from the Twitter stream using the detection algorithm described in [8]. The detection algorithm looks for Tweets which contain the word "earthquake" and computes short-term-average (one-minute) divided by long-term-average (sixty minutes). The authors state that their formula is commonly used in seismology to detect and time seismic phases. The algorithm includes user-set parameters which can be used to tune the sensitivity of the algorithm. We tested their algorithm using their recommended parameters for sensitive, moderate, and conservative detection using the "earthquake" keyword as well as additional keywords associated with earthquakes, such as "shaking". The selected keywords were determined by looking at the Tweets following a magnitude 5 earthquake in Southern California and using Google Correlate[15] [50].

The authors of [8] state that as the sensitivity is increased (i.e., as it becomes less conservative), more false positives will be reported. As expected, we found that as we increased sensitivity more earthquakes were reported, but most of these reports did correlate with a known earthquake nor did they seem to identify an actual unknown earthquake (i.e., an earthquake which was not detected by USGS). The results of running the algorithm are summarized in Table 3.3.

[15] https://www.google.com/trends/correlate

Table 3.3. The number of earthquakes reported by [8] on our data for various settings

Detection Setting	"Earthquake" Only	Additional Keywords
Conservative	5	5
Moderate	27	35
Sensitive	100	164

Next, we attempted to detect earthquakes using text analysis and machine learning techniques. We utilized Sandia's Cognitive Foundry library, which is a robust software library for Cognitive Science and Technology applications [21] that provides an extensive library of powerful machine learning algorithms.

We trained machine learning algorithms with approximately 5,000 manually classified Tweets from the keywords database. The features we examined included: (1) the length of the tweet, (2) the number of unique keywords in the tweet, and (3) the total number of keywords in the tweet. This approach is similar to the approach in [25].Unfortunately, we were not able to develop a model which was able to correctly determine whether or not a tweet referred to an actual earthquake. This is due, in part, to the short length of tweets.

Soft data may be a useful tool for making predictions, but as the authors of [43] observed, it should not be a standalone alone tool. Rather is should be combined with hard data in order make more accurate predictions. Finding useful information about interesting events using social media can be like finding a needle in a haystack.

4. CONCLUSION

DF is an important problem which has many useful applications, but there is still more work to be done in order to fully exploit the information contained in the vast amounts of available data. This project has brought many of the available tools together in order to start exploiting the available information

This page intentionally left blank

REFERENCES

1. Big data at the speed of business. http://www-01.ibm.com/software/data/bigdata/ what-is-big-data.html. Accessed: May 2015.
2. Alessandro Zanasi. Virtual Weapons for Real Wars: Text Mining for National Security, pages 53–60. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
3. William C. Lisse. The economics of information and the internet. Competitive Intelligence Review, 9(4):48–55, 1998.
4. T. Abirami, E. Taghavi, R. Tharmarasa, T. Kirubarajan, and A. C. Boury-Brisset. Fusing social network data with hard data. In 2015 18th International Conference on Information Fusion (Fusion), pages 652–658, July 2015.
5. Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. CoRR, abs/1704.00656, 2017.
6. Zeynep Tufekci and Christopher Wilson. Social media and the decision to participate in political protest: Observations from tahrir square. Journal of Communication, 62(2):363–379, 2012.
7. Why “big data” is a big deal: Information science promises to change the world. Harvard Magazine, 2014. Accessed: April 2015.
8. Paul Earle, Daniel Bowden, and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. Annals of Geophysics, 54(6), 2012.
9. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1):97–107, January 2014.
10. Jens Bleiholder and Felix Naumann. Data fusion. ACM Computing Surveys, 41(1):1:1–1:41, January 2009.
11. Xin Luna Dong and Divesh Srivastava. Big data integration. In 2013 IEEE 29th International Conference on Data Engineering (ICDE), pages 1245–1248, April 2013.
12. Martin Wainwrite. New theory and algorithms for scalable data fusion. Technical report, UC Berkely and Air Force Research Laboratory, July 2013.
13. Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: Is the problem solved? In Proceedings of the VLDB Endowment, volume 6, pages 97–108, December 2012.
14. Bahador Khaleghi and Fakhreddine Karray. Random set theoretic soft/hard data fusion framework. IEEE Transactions on Aerospace and Electronic Systems, 50(4):3068–3081, October 2014.
15. Waldo R. Tobler. A computer moview simulating urban growth in the detroit region. Economic Geography, 46(2):234–240, June 1970.
16. Ceren Budak, Theodore Georgiou, Divyakant Agrawal, and Amr El Abbadi. Geoscope: Online detection of geo-correlated information trends in social networks. Proc. VLDB Endow., 7(4):229–240, December 2013.

17. Seung Woo Cho, MoonSu Cha, and Kyung-Ah Sohn. Topic category analysis on twitter via cross-media strategy. Multimedia Tools and Applications, 75(20):12879–12899, June 2016.

18. Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In Association for Computational Linguistics (ACL) System Demonstrations, pages 55– 60, 2014.

19. Renxian Zhang, Wenjie Li, Dehong Gao, and You Ouyang. Automatic twitter topic summarization with speech acts. IEEE Transactions on Audio, Speech, and Language Processing, 21(3):649–658, March 2013.

20. Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Dynamic theme tracking in twitter. In 2015 IEEE International Conference on Big Data (Big Data), pages 561–570, October 2015.

21. Justin Basilico, Zachary Benz, and Kevin Dixon. The cognitive foundry: A flexible platform for intelligent agent modeling. In Proceedings of the 2008 Behavior Representation in Modeling and Simulation Conference, pages 61–70, April 2008.

22. Fabio Ciulla, Delia Mocanu, Andrea Baronchelli, Bruno Gon¸calves, Nicola Perra, and Alessandro Vespignani. Beating the news using social media: the case study of american idol. EPJ Data Science, 1(1):8, July 2012.

23. Hyunyoung Choi and Hal Varian. Predicting the present with google trends. Economic Record, 88:2–9, 2012.

24. L Burks, M Miller, and R Zadeh. Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets. In NCEE 2014 - 10th U.S. National Conference on Earthquake Engineering: Frontiers of Earthquake Engineering, January 2014.

25. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

26. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Transactions on Knowledge and Data Engineering, 25(4):919–931, April 2013.

27. Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

28. Adam Bermingham and Alan Smeaton. On using twitter to monitor political sentiment and predict election results. In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 2–10, Chiang Mai, Thailand, November 2011. IJCNLP.

29. Johen Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2(1):1–8, March 2011.

30. Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown. Automatic Crime Prediction Using Events Extracted from Twitter Posts, pages 231–238. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

31. Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, pages 78–87, New York, NY, USA, 2005. ACM.

32. Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Arsa: A sentiment-aware model for predicting sales performance using blogs. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, pages 607–614, New York, NY, USA, 2007. ACM.

33. Gilad Mishne and Natalie Glance. Predicting movie sales from blogger sentiment. In Proceedings ofAAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs, Stanford, US, January 2006.

34. Sharad Goel, Jake M. Hofman, S´ebastien Lahaie, David M. Pennock, and Duncan J. Watts. Predicting consumer behavior with web search. Proceedings of the National Academy of Sciences, 107(41):17486–17490, 2010.

35. Robert P. Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. ACM Trans. Inf. Syst., 27(2):12:1–12:19, March 2009.

36. Patrick Copeland, Raquel Romano, Tom Zhang, Greg Hecht, Dan Zigmond, and Christian Stefansen. Google disease trends: An update. In International Society of Neglected Tropical Diseases 2013, page 3, 2013.

37. Declan Butler. When google got flu wrong. Nature, 494:155–156, February 2013.

38. Sarah Milstein, Abdur Chowdhury, Gregor Hochmuth, Ben Lorica, Roger Magoulas, and Tim O'Reilly. Twitter and the Micro-Messaging Revolution, chapter Introduction: What is Twitter?, pages 3–21. O'Reilly, November 2008.

39. Twitter. https://en.wikipedia.org/wiki/Twitter. Accessed: August 2017.

40. Number of monthly active twitter users worldwide from 1st quarter 2010 to 2nd quarter 2017 (in millions). https://www.statista.com/statistics/282087/ number-of-monthly-active-twitter-users/. Accessed: August 2017.

41. Cooper Smith. These are the most twitter-crazy countries in the world, starting with saudi arabia(!?). http://www.businessinsider.com/ the-top-twitter-markets-in-the-world-2013-11, November 2013. Accessed: August 2017.

42. The one million tweet map. http://onemilliontweetmap.com/. Accessed: August 2017.

43. David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: Traps in big data analysis. Science, 343(6176):1203–1205, 2014.

44. Samantha Finn, Panagiotis Takis Metaxas, and Eni Mustafaraj. Spread and skepticism: Metrics of propagation on twitter. In Proceedings of the ACM Web Science Conference, page 39, June 2015.

45. Pacific Northwest Seismic Network. Magnitude/intensity. https://pnsn.org/outreach/about-earthquakes/magnitude-intensity. Accessed: August 2017.

46. Usgs. https://www.usgs.gov/about/about-us. Accessed: August 2017.

47. Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, and Jiawei Han. Pet: A statistical model for popular events tracking in social communities. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, pages 929–938, New York, NY, USA, 2010. ACM.

48. Cindy Xide Lin, Qiaozhu Mei, Jiawai Han, Yunliang Jiang, and Marina Danilevsky. The joint inference of topic diffusion and evolution in social communities. In 2011 IEEE 11th International Conference on Data Mining, pages 378–387, December 2011.

49. Usgs maps. https://earthquake.usgs.gov/earthquakes/map. Accessed: August 2017.

50. Matt Mohebbi, Dan Vanderkam, Julia Kodysh, Rob Schonberger, Hyunyoung Choi, and Sanjiv Kumar. Google correlate whitepaper. Technical report, June 2011.

This page intentionally left blank

DISTRIBUTION

1	MS0899	Technical Library	9536 (electronic copy)
1	MS0359	D. Chavez, LDRD Office	1911


Sandia National Laboratories